Refined Myocardium Segmentation from CT Using a Hybrid-Fusion Transformer

Shihua Qin,* Fangxu Xing,* Jihoon Cho, Jinah Park, *Senior Member, IEEE*, Xiaofeng Liu, Amir Rouhollahi, Elias J. Bou Farhat, Hoda Javadikasgari, Ashraf Sabe, Farhad R. Nezami, *Member, IEEE*, Jonghye Woo,# *Senior Member, IEEE*, and Iman Aganj#

***Abstract*—Accurate segmentation of the left ventricle (LV) in cardiac CT images is crucial for assessing ventricular function and diagnosing cardiovascular diseases. Common semi-automatic segmentation often includes unwanted structures, such as papillary muscles, due to low contrast between the LV wall and surrounding tissues. In this study, we address this issue by proposing a two-input-channel method within a Hybrid-Fusion Transformer deep-learning framework. Our method refines coarse LV masks by incorporating both the CT images and the semi-automatic rough masks as input channels, effectively removing papillary muscles. Using a small number of manually refined labels, we evaluated the method through leave-one-out cross-validation. The results demonstrate an average Dice similarity coefficient of 95.2%, outperforming models that use only CT images or rough masks. This approach reduces the need for extensive manual labeling while maintaining high segmentation accuracy and stability, making it suitable for clinical applications.**

***Index Terms*—Cardiac imaging, myocardium, papillary muscle, segmentation, CT, deep learning, transformer, refinement.**

I. Introduction

Early detection of morphological and functional abnormalities of the myocardium is essential for preventing cardiovascular diseases [1], which remain a leading cause of mortality worldwide [2]. In cardiology, ventricular performance is typically assessed through measurements such as wall thickness, ventricular volume, and ejection fraction [3]. These critical indices are often obtained through the segmentation of cardiac images produced by medical imaging modalities such as magnetic resonance imaging and computed tomography (CT) [4].

This paper was initially submitted for review on September 25, 2024. This work was supported by the Connors Center for Women's Health and Gender Biology at the Brigham and Women's Hospital (BWH) and the Office of Chief Academic Officer at Mass General Brigham (MGB) through a Connors BWH-MGB Collaborative IGNITE Award.

Shihua Qin (kishoreq@bu.edu) is with Questrom School of Business, Boston University, and the Department of Radiology, Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA.

Fangxu Xing (fxing1@mgh.harvard.edu), Jonghye Woo, and Iman Aganj are with the Department of Radiology, Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA.

Jihoon Cho and Jinah Park are with the School of Computing, Korea Advanced Institute of Science and Technology, Daejeon, South Korea.

Xiaofeng Liu is with the Department of Radiology and Biomedical Imaging, Yale University, New Haven, CT, USA.

Amir Rouhollahi, Hoda Javadikasgari, and Farhad Nezami are with the Department of Surgery, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA.

Elias Bou Farhat and Ashraf Sabe are with the Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA.

*S. Qin and F. Xing contributed equally to this work.

#J. Woo and I. Aganj are co-senior authors.

In recent years, deep neural networks have greatly aided in improving the performance of medical image segmentation across different modalities [5]. Convolutional neural network (CNN)-based learning models, such as the U-Net [6], are commonly adopted in segmentation [7]. Through its skip connections, U-Net allows high-level features in the shallow layers to propagate to the higher-resolution layers in the decoder. Chen et al. proposed an improved CNN-based model incorporating both motion estimation and segmentation, achieving a high Dice similarity coefficient (DSC) of 0.93 [8]. Isensee et al. utilized an ensemble of U-Net-inspired architectures for segmenting multiple cardiac structures including the left ventricle (LV) cavity, demonstrating CNN's superior performance in myocardium segmentation [9]. These models, however, typically require a large amount of pre-labeled images, usually with manually delineated masks, to use as gold-standard for training, which is time-consuming and labor-intensive to obtain, especially for 3D volumes. Therefore, cardiologists resort to semi-automatic labeling techniques to facilitate training data delineation [10], [11]. These methods are often intensity-based and therefore not sufficiently robust to separate smaller image details from their adjacent structures [12]. Specifically, due to the low contrast between the LV wall and papillary muscles in CT images, semi-automatically generated LV masks largely include adjacent papillary muscles. Several previous studies have shown that exclusion of papillary muscles yields the most reproducible endocardial segmentation [13], [14]. Additionally, a smooth endocardium boundary can ensure more accurate measurements of the ejection fraction and the volume [15]. Thus, proper refinement of the semi-automatic coarse masks – which cannot be practically achieved by resource-demanding manual delineations – is needed to create papillary-muscle-free input labels for deep-network training.

In this work, we propose a transformer-based CNN model for the purpose of segmentation refinement and papillary muscle removal from CT images of the myocardium [16]. The model is trained on a small dataset of subjects, each with a CT image, a semi-automatic coarse mask, and a manually refined mask, whereby it takes both the image and the coarse mask as two-channel input and learns to produce the refined mask. The CNN part of the model handles the coarse segmentation and the transformer part handles the refining of the coarse masks to exclude papillary muscles [17]. By leveraging information from a small amount of fine labels, the proposed model allows us to fully exploit the efficiency of semi-automatic labeling, while mitigating the high cost of extensive manual

refinements. Our evaluation shows improved segmentation accuracy compared to using only the single-channel input.

II. Method

For each training CT image I, we denote its semi-automatic coarse LV mask containing papillary muscles as M^c and its manually refined mask without the papillary muscles as M^r. The proposed network focuses on leveraging information from both I and M^c to predict a refined, smooth LV region, given a limited availability of cases with M^r.

The deep-learning model we propose is based on the Hybrid-Fusion Transformer (HFTrans) architecture [16]. HFTrans combines the strengths of both CNNs and transformers, where the CNN encoder is used to extract contextual features from each input image-and-label pair, and the transformer layers exert self-attention on these extracted features to enhance representation and connect information across channels. Based on this framework, we propose to expand the input image to two channels defined in $\mathbb{R}^{2\times W\times H\times D}$ (width, height, depth). For the ith training image, channel 1 is the original cardiac CT volume $I_i \in \mathbb{R}^{1\times W\times H\times D}$ and channel 2 is its corresponding coarse mask $M_i^c \in \mathbb{R}^{1\times W\times H\times D}$. The stacked new input $\tilde{x}_i$= $[I_i, M_i^c]$ provides a richer context, where the coarse mask captures the overall structure of the LV (where the refined mask should be confined to) and its surrounding anatomy, and the CT image provides voxel-level contextual information. Correspondingly, the manually refined $M_i^r \in \mathbb{R}^{1\times W\times H\times D}$ is used as the input training label for image i. This augmented input allows the transformer part of the network to focus on correcting regional details and producing a refined mask.

For each input pair $\tilde{x}_i$, the features f_j (where $j = 1, 2, 3$) are extracted from three CNN encoders. The first encoder performs early fusion by processing both the CT image and rough mask together, while the other two encoders extract features individually from each sequence (I_i or M_i^c). These features are then linearly projected, and a positional embedding is added to retain spatial information, as shown in Fig 1. The initial embedding z_0 is calculated as:

$$z_0 = W \times f + E_{pos},$$

where W is the projection matrix and E_{pos} is the positional embedding. The Transformer then applies a multi-head self-attention (MSA) and a multilayer perceptron (MLP), with layer normalization (LN), to refine the features across multiple layers, with each Transformer layer defined as:

$$z_l^* = MSA(LN(z_{l-1})) + z_{l-1}, \quad z_l = MLP(LN(z_l^*)) + z_l^*.$$

To deal with limited input data with refined labels, we applied several data augmentation techniques. These include flipping, rotation, and scaling applied to both the CT images and their corresponding masks. Additionally, due to GPU memory limitations, the original input size of $160 \times 160 \times 160$ was cropped to $128 \times 128 \times 128$ at a random location. In the pre-processing stage, we exclusively performed contrast adjustments, Gaussian noise addition, and histogram equalization on the CT images. For post-processing, we enhanced model stability and performance by flipping the input patch along all three dimensions, generating 8 different spatial perspectives of the same object. We then predicted 8 probability maps, and after flipping them back to their original orientations, we averaged them to create the final probability map, improving both performance and stability.

III. Experimental Results and Discussion

A. Data

1) Data Description: We retrospectively identified a dataset of 35 adult patients with severe aortic stenosis who underwent transcatheter aortic valve replacement (TAVR) at Brigham and Women's Hospital (Boston, MA, USA) from 2018 to 2020 [18]. We excluded patients under 18 years of age, as well as those with mixed aortic regurgitation, more than mild concomitant valvular disease, a history of ischemic heart disease, prior valve or coronary artery bypass graft surgery, or poor-quality pre-TAVR gated cardiac CT images. CT images in the cardiac region were retrospectively collected in the DICOM format with approval from the local institutional review board [19]. The raw DICOM data were initially de-identified by an independent party, converting the images into the nearly raw raster data (NRRD) format. Images were downsampled to the size $160 \times 160 \times 160$, resulting in an in-plane pixel size varying from 0.89 to 1.44 mm, and a slice thickness varying from 0.69 to 1.15 mm.

2) Semi-automatic mask creation: A trained expert used the open-source image processing software 3D Slicer [20] to semi-automatically segment the LV using the "grow from seed" option, which employs a conventional region-growing-based method using sample points (seeds) for the foreground (LV) and background in selected slices chosen by the clinician to classify the myocardial regions in both the selected and remaining slices. After the initial iteration, the slices were reviewed to identify misclassified regions, and appropriate seeds were added to improve the mask. This iterative process continued until the desired segmentation was achieved. Finally, the segmented masks were converted to a binary label file.

3) Manual mask refinement: Since the presence of papillary muscles within the semi-automatic segmented labels was not desired, we selected five cases where these muscles were distinguishable by a trained clinician. Our whole dataset encompassed a range of papillary muscle quality and appearance, e.g. some images exhibited lower contrast and more noise, and the five cases were specifically chosen to exclude these imperfections. Manual removal of the papillary muscles was performed using the erasing tool on 3D Slicer for each slice and each case by a clinician.

B. Validation scheme

We adopted two key metrics: the DSC and the 95% Hausdorff Distance (HD95) [21]. The DSC evaluates the overlap between the predicted and ground-truth masks, providing a measure of segmentation accuracy. The HD95, on the other hand, assesses the boundary alignment, ensuring that the refined segmentation is not only accurate but also spatially consistent with the ground truth.

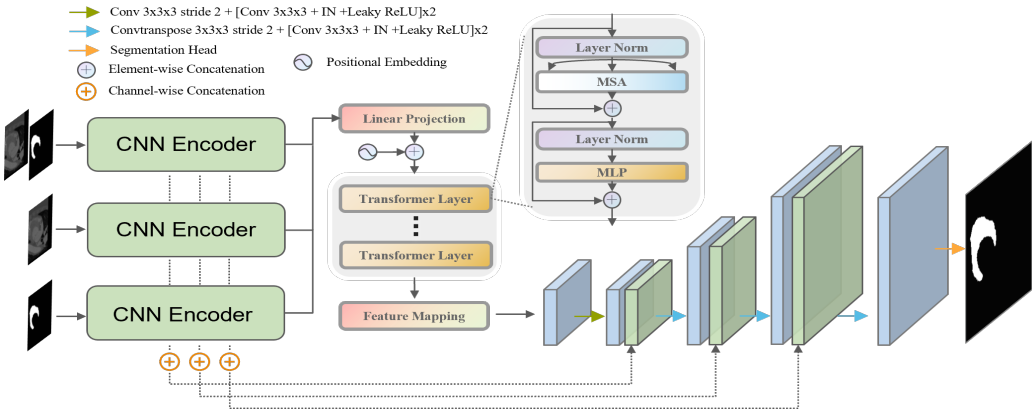

Fig. 1. Overview of our neural network structure, which consists of three CNN encoders. The first encoder performs early fusion of CT images and the rough mask, while the other two encoders are used to extract contextual feature representations from the CT image and rough mask separately. A Transformer encoder is then used to integrate the features extracted by these encoders.

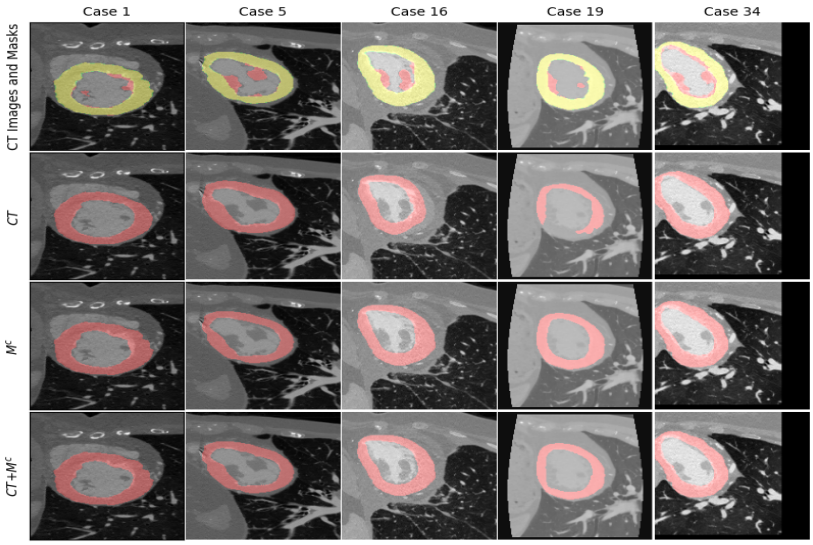

Fig. 2. Qualitative comparison of different input types across the 5 cases. In the first row, the manual yellow mask represents the LV wall excluding the papillary muscle, while the semi-automatic red mask indicates the unwanted papillary muscle. The remaining three rows show the model-predicted LV segmentation using different input types: CT image only, rough mask only, and the combination of CT image and rough mask, respectively.

To determine the contribution of each channel of $\tilde{x}_i$ on the performance of the model, we designed and tested three variations: using the original CT image $\tilde{x}_{i,1} = x_i$, using the predicted rough mask alone $\tilde{x}_{i,2} = M^c$, and using the combined stack of the CT image and rough mask $\tilde{x}_{i,3} = \text{stack}(x_i, M^c)$.

Given the limited availability of data, with only five cases, we employed a cross-validation strategy to maximize the utility of our dataset. Specifically, we used a leave-one-out cross-validation approach, where four cases were used for training and the remaining case was reserved for validation. This approach ensures that every data point is used both for training and validation, providing a robust estimate of the model's performance.

TABLE I
COMPARISON OF THREE DIFFERENT INPUT STRATEGIES

	CT		M^c		$CT + M^c$	
	DSC↑ (%)	**HD95↓** (voxels)	**DSC↑** (%)	**HD95↓** (voxels)	**DSC↑** (%)	**HD95↓** (voxels)
Case 1	91.2	2.45	94.8	2.85	**96.5**	**1.46**
Case 5	87.6	3.16	**96.7**	**1.23**	96.6	1.51
Case 16	80.6	7.87	**95.8**	**2.24**	95.0	2.74
Case 19	71.4	10.68	**97.2**	**1.24**	93.4	3.34
Case 34	89.2	3.61	92.4	3.35	**94.6**	**2.24**
Mean	84.0	5.55	**95.4**	**2.18**	95.2	2.25
StD	8.1	3.56	1.9	0.95	**1.4**	**0.81**

C. Quantitative results

The results are presented in Table I and Fig. 2. When using only CT images as input, the model struggled to achieve satisfactory performance, possibly due to limited training data. The average DSC was 84.0%, with a high standard deviation for both DSC (8.1) and HD95 (3.56). Notably, Case 19 exhibits a significantly different pixel distribution compared to the

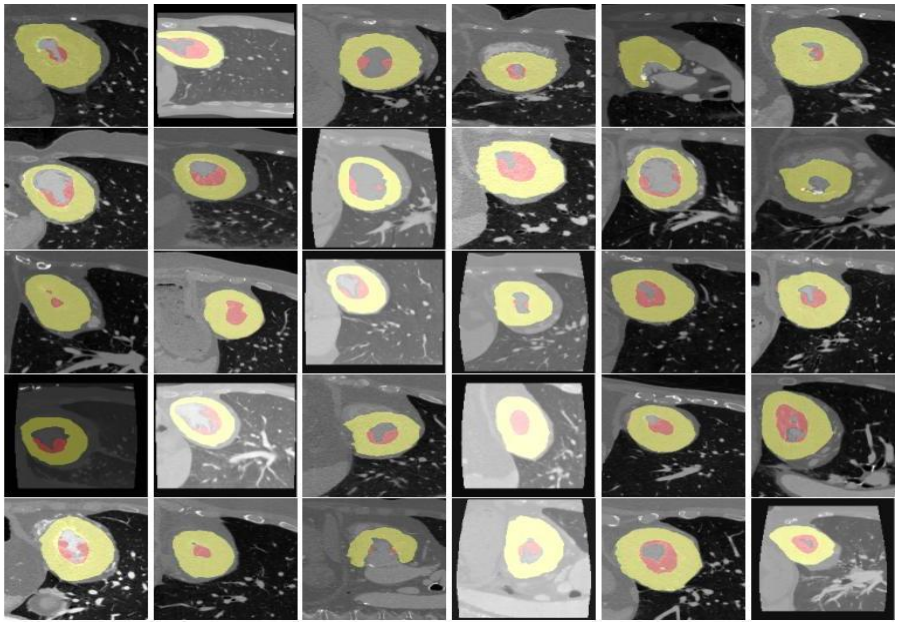

Fig. 3. Refined segmentation predictions for LV walls of the remaining 30 adult patients are presented. Each patient's dataset includes a semi-automatically generated mask, with the papillary muscles highlighted in red.

other cases, resulting in a DSC of only 71.4%. This occurred despite the application of data augmentation techniques during training to enhance image diversity. Additionally, for Case 19, we applied pre-processing techniques, including histogram matching (a function in the Python skimage package), to align the pixel intensity distribution of the input image with that of a reference image (Case 1).

Second, it is interesting to observe that the two strategies—one using only the mask with papillary muscles (M^c) as input and the other with two-channel inputs (CT + M^c)—achieved better results than using only CT images. The M^c-only input strategy resulted in an average DSC of 95.4% and a significantly lower standard deviation (1.9 for DSC and 0.95 for HD95), indicating that the model can generalize better across different cases when using the M^c mask. This suggests that the M^c mask, which provides the position and rough contours of the LV, contains most of the necessary information to predict the smooth mask accurately. Interestingly, the combination of CT and M^c inputs led to a close average DSC compared to using only M^c inputs (95.2% vs. 95.4%). However, the combined input strategy achieved the lowest standard deviation for both DSC (1.35) and HD95 (0.81), demonstrating that the additional CT image data helps to stabilize the predictions. This underscores the importance of incorporating both anatomical structure and microenvironment information in the input data.

In certain cases, such as Case 19, the variability introduced by the CT image—likely due to differences in pixel histogram distributions—negatively affected the performance of the two-channel method. For this case, the DSC for the two-channel input was 93.4%, compared to 97.2% with the single-M^c-channel input, a much larger difference than observed in other cases. This suggests that while the CT image adds valuable contextual information, it can also introduce noise in cases with significant image intensity variation, particularly in smaller datasets. Despite these challenges, the two-channel method remains essential as it provides a more comprehensive understanding of the anatomical context.

D. Applicability

To evaluate the effectiveness of our method in removing the papillary muscle, we present in Fig 3 the refined LV wall predictions for the remaining 30 subjects, using a two-channel model trained on the 5 subjects with manually refined masks. The results demonstrate that, despite being trained on a small dataset, the model successfully recognizes and excludes the papillary muscle, proving its practicality in real-world applications. In particular, the resulting refined labels could be used as silver-standard labels to train a single-CT-input network on all 35 images, enabling refined LV segmentation of a larger CT database (for which coarse labels are unavailable).

IV. Conclusion

In this paper, we presented a two-channel Hybrid-Fusion Transformer-based deep-learning method for refining semi-automatic LV segmentation in cardiac CT images, specifically focused on excluding the papillary muscles. By utilizing both CT images and rough masks as input, the method integrates complementary information to enhance segmentation accuracy and stability. Although the two-channel method did not consistently outperform the single-channel rough-mask input in all cases, the inclusion of the CT image remains valuable. It provides essential anatomical context around the LV, helping the model to refine segmentation where the rough mask alone may miss critical details. By combining the structural information from rough masks with the contextual data from CT images, the method generally yields more reliable and consistent segmentation results in most cases.

REFERENCES

[1] W. H. Gaasch and M. R. Zile, "Left ventricular structural remodeling in health and disease," *Journal of the American College of Cardiology*, vol. 58, no. 17, pp. 1733–1740, 2011. [Online]. Available: https://www.jacc.org/doi/abs/10.1016/j.jacc.2011.07.022

[2] G. A. Roth, G. A. Mensah, and V. Fuster, "The global burden of cardiovascular diseases and risks: a compass for global action," pp. 2980–2981, 2020.

[3] M. A. Fifer and W. Grossman, "Measurement of ventricular volumes, ejection fraction, mass, wall stress, and regional wall motion," *Cardiac catheterization, angiography and intervention. 4th ed. Philadelphia: Lea & Febiger*, pp. 310–312, 1991.

[4] P. Peng, K. Lekadir, A. Gooya, L. Shao, S. E. Petersen, and A. F. Frangi, "A review of heart chamber segmentation for structural and functional analysis using cardiac magnetic resonance imaging," *Magnetic Resonance Materials in Physics, Biology and Medicine*, vol. 29, pp. 155–195, 2016.

[5] X. Liu, L. Song, S. Liu, and Y. Zhang, "A review of deep-learning-based medical image segmentation methods," *Sustainability*, vol. 13, no. 3, p. 1224, 2021.

[6] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3D U-Net: learning dense volumetric segmentation from sparse annotation," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2016: 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part II 19*. Springer, 2016, pp. 424–432.

[7] A. Rouhollahi, J. N. Willi, S. Haltmeier, A. Mehrtash, R. Straughan, H. Javadikasgari, J. Brown, A. Itoh, K. I. de la Cruz, E. Aikawa, E. R. Edelman, and F. R. Nezami, "Cardiovision: A fully automated deep learning package for medical image segmentation and reconstruction generating digital twins for patients with aortic stenosis," *Computerized Medical Imaging and Graphics*, vol. 109, p. 102289, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0895611123001076

[8] C. Qin, W. Bai, J. Schlemper, S. E. Petersen, S. K. Piechnik, S. Neubauer, and D. Rueckert, "Joint learning of motion estimation and segmentation for cardiac mr image sequences," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part II 11*. Springer, 2018, pp. 472–480.

[9] F. Isensee, P. F. Jaeger, P. M. Full, I. Wolf, S. Engelhardt, and K. H. Maier-Hein, "Automatic cardiac disease assessment on cine-mri via time-series segmentation and domain specific features," in *Statistical Atlases and Computational Models of the Heart. ACDC and MMWHS Challenges: 8th International Workshop, STACOM 2017, Held in Conjunction with MICCAI 2017, Quebec City, Canada, September 10-14, 2017, Revised Selected Papers 8*. Springer, 2018, pp. 120–129.

[10] J. Egger, T. Kapur, A. Fedorov, S. Pieper, J. V. Miller, H. Veeraraghavan, B. Freisleben, A. J. Golby, C. Nimsky, and R. Kikinis, "Gbm volumetry using the 3D Slicer medical image computing platform," *Scientific reports*, vol. 3, no. 1, p. 1364, 2013.

[11] E. R. Velazquez, C. Parmar, M. Jermoumi, R. H. Mak, A. Van Baardwijk, F. M. Fennessy, J. H. Lewis, D. De Ruysscher, R. Kikinis, P. Lambin *et al.*, "Volumetric CT-based segmentation of NSCLC using 3D-Slicer," *Scientific reports*, vol. 3, no. 1, p. 3529, 2013.

[12] J. Fan, G. Zeng, M. Body, and M.-S. Hacid, "Seeded region growing: an extensive and comparative study," *Pattern recognition letters*, vol. 26, no. 8, pp. 1139–1156, 2005.

[13] K. Jaspers, H. G. Freling, K. Van Wijk, E. I. Romijn, M. J. Greuter, and T. P. Willems, "Improving the reproducibility of mr-derived left ventricular volume and function measurements with a semi-automatic threshold-based segmentation algorithm," *The international journal of cardiovascular imaging*, vol. 29, pp. 617–623, 2013.

[14] Y. Liu, G. Captur, J. C. Moon, S. Guo, X. Yang, S. Zhang, and C. Li, "Distance regularized two level sets for segmentation of left and right ventricles from cine-mri," *Magnetic resonance imaging*, vol. 34, no. 5, pp. 699–706, 2016.

[15] L. Xie, Y. Song, and Q. Chen, "Automatic left ventricle segmentation in short-axis MRI using deep convolutional neural networks and central-line guided level set approach," *Computers in Biology and Medicine*, vol. 122, p. 103877, 2020.

[16] J. Cho and J. Park, "Hybrid-fusion transformer for multisequence MRI," in *International Conference on Medical Imaging and Computer-Aided Diagnosis*. Springer, 2022, pp. 477–487.

[17] R. Strudel, R. Garcia, I. Laptev, and C. Schmid, "Segmenter: Transformer for semantic segmentation," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 7262–7272.

[18] R. R. Makkar, G. P. Fontana, H. Jilaihawi, S. Kapadia, A. D. Pichard, P. S. Douglas, V. H. Thourani, V. C. Babaliaros, J. G. Webb, H. C. Herrmann *et al.*, "Transcatheter aortic-valve replacement for inoperable severe aortic stenosis," *New England Journal of Medicine*, vol. 366, no. 18, pp. 1696–1704, 2012.

[19] M. Onken, M. Eichelberg, J. Riesmeier, and P. Jensch, "Digital imaging and communications in medicine," in *Biomedical Image Processing*. Springer, 2010, pp. 427–454.

[20] S. Pieper, M. Halle, and R. Kikinis, "3D Slicer," in *2004 2nd IEEE international symposium on biomedical imaging: nano to macro (IEEE Cat No. 04EX821)*. IEEE, 2004, pp. 632–635.

[21] J. Ma, J. Chen, M. Ng, R. Huang, Y. Li, C. Li, X. Yang, and A. L. Martel, "Loss odyssey in medical image segmentation," *Medical Image Analysis*, vol. 71, p. 102035, 2021.